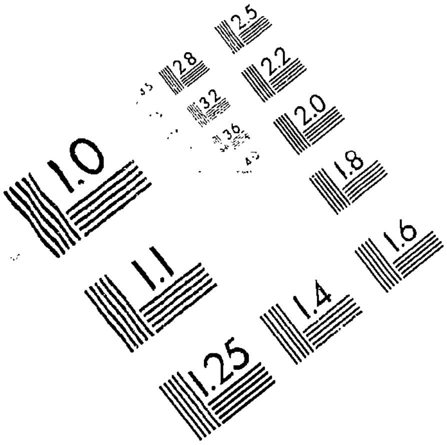
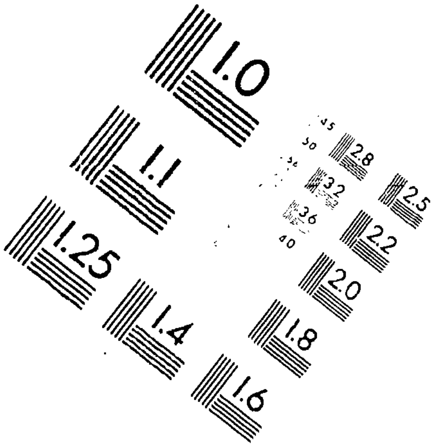
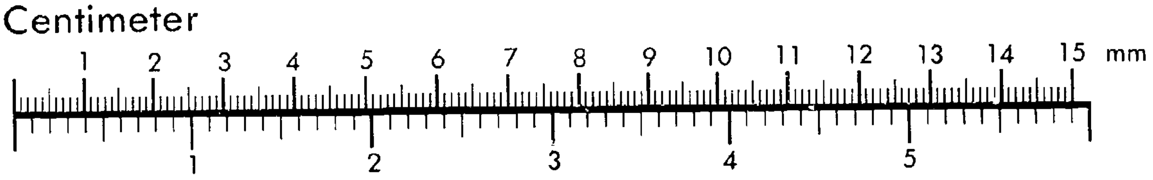


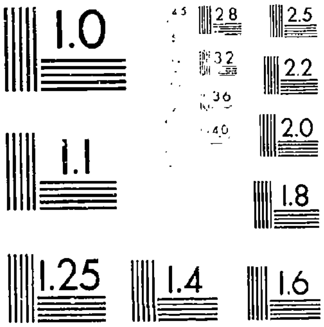
AIIM

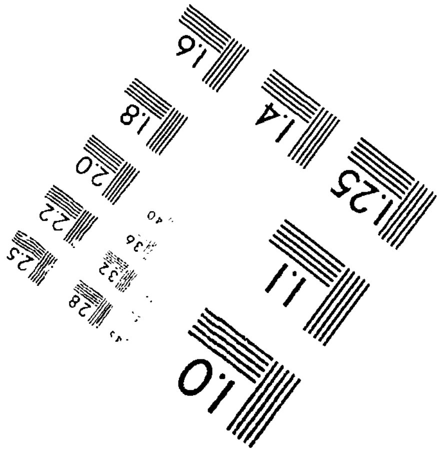
Association for Information and Image Management

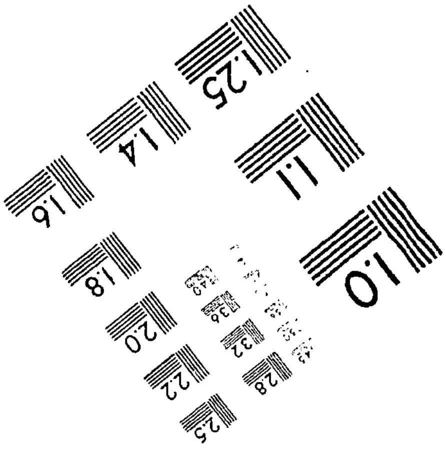
1100 Wayne Avenue, Suite 1100
Silver Spring, Maryland 20910
301/587-8202











MANUFACTURED TO AIIM STANDARDS
BY APPLIED IMAGE, INC.

ED 357 086 TM 019 865

AUTHOR	Kim, Seock-Ho; And Others
TITLE	Detection of Differential Item Functioning in Multiple Groups.
PUB DATE	Apr 93
NOTE	33p.; Paper presented at the Annual Meeting of the American Educational Research Association (Atlanta, GA, April 12-16, 1993).
PUB TYPE	Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	*Chi Square; College Students; *Comparative Analysis; Equations (Mathematics); *Estimation (Mathematics); *Groups; Higher Education; *Item Bias; Item Response Theory; *Mathematical Models; Research Methodology; Sample Size
IDENTIFIERS	Item Parameters; Q Statistic

ABSTRACT

Detection of differential item functioning (DIF) is most often done between two groups of examinees under item response theory. It is sometimes important, however, to determine whether DIF is present in more than two groups. A method is presented for the detection of DIF in multiple groups. The method, the Q(sub j) statistic, is closely related to the chi-square method of F. M. Lord for comparing vectors of item parameters estimated in two groups and is based on the same assumptions. An example is provided using real data from 3 groups of 200 college students each (1 reference and 2 focal groups). Five tables present analysis results, and a 49-item list of references is included. (Author/SLD)

DOCUMENT RESUME

ED 357 086 TM 019 865

AUTHOR Kim, Seock-Ho; And Others

TITLE Detection of Differential Item Functioning in Multiple Groups.

PUB DATE Apr 93

NOTE 33p.; Paper presented at the Annual Meeting of the American Educational Research Association (Atlanta, GA, April 12-16, 1993).

PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.

DESCRIPTORS *Chi Square; College Students; *Comparative Analysis; Equations (Mathematics); *Estimation (Mathematics); *Groups; Higher Education; *Item Bias; Item Response Theory; *Mathematical Models; Research Methodology; Sample Size

IDENTIFIERS Item Parameters; Q Statistic

ABSTRACT

Detection of differential item functioning (DIF) is most often done between two groups of examinees under item response theory. It is sometimes important, however, to determine whether DIF is present in more than two groups. A method is presented for the detection of DIF in multiple groups. The method, the Q(sub j) statistic, is closely related to the chi-square method of F. M. Lord for comparing vectors of item parameters estimated in two groups and is based on the same assumptions. An example is provided using real data from 3 groups of 200 college students each (1 reference and 2 focal groups). Five tables present analysis results, and a 49-item list of references is included. (Author/SLD)

Detection of Differential Item Functioning in Multiple Groups

Seock-Ho Kim, Allan S. Cohen, and Tae-Hak Park
University of Wisconsin–Madison

Running Head: DIF IN MULTIPLE GROUPS

April, 1993
Paper presented at the annual meeting of
the American Educational Research Association
at Atlanta, Georgia.

Detection of Differential Item Functioning in Multiple Groups

Abstract

Detection of differential item functioning (DIF) is most often done between two groups of examinees under item response theory. It is sometimes important, however, to determine whether DIF is present in more than two groups. In this paper we present a method for detection of DIF in multiple groups. The method is closely related to Lord's chi-square for comparing vectors of item parameters estimated in two groups. An example using real data is provided.

Key words: differential item functioning, item response theory, Lord's chi-square.

Introduction

An item is said to be di.rentially functioning if the probability of a correct response is different for examinees at the same ability level but from different groups (cf. Pine, 1977). Efforts to detect differential item functioning (DIF) have been extensively reviewed by Berk (1982) and Holland and Wainer (1993) for methods based on both classical test theory and item response theory (IRT). DIF detection methods under either theoretical approach, however, have been most completely developed for the two group case in which comparisons are made between some aspect of the responses of examinees in a base (or reference) group and examinees in a second (or focal) groups. It is not uncommon, however, to have a situation in which more than two groups exist. With most current DIF detection methods, multiple two-group comparisons are required to detect LIF across all groups. This approach does not permit other than pairwise comparisons among groups. A more appropriate and useful approach would be to search for DIF simultaneously across all groups. In this paper, we present a method under IRT for simultaneous detection of DIF in a multiple group situation.

Detection of DIF under IRT is based on the assumption that the items on the test measure the same underlying ability in all groups from the same population. Two main approaches have been used for detection of DIF under IRT. One approach focuses on a comparison of item parameters estimated in two groups (Draba, 1977; Lord, 1977, 1980; Thissen, Steinberg, & Wainer, 1988, 1993; Wright & Stone, 1979). The other approach focuses on the area between item response functions (IRFs) from two groups (Kim & Cohen, 1991; Linn, Levine, Hastings, & Wardrop, 1981; Raju, 1988, 1990; Rudner, 1977; Wainer, 1993). Our focus in this paper is on the first set of methods which compare item parameters estimated in different groups. Specifically, we describe a chi-square method for comparison of item parameters estimated in multiple groups.

IRT Model

The probability of a correct response for a dichotomously scored item can be expressed by the three-parameter IRF (Birnbaum, 1968) as

$$P_j(\theta) = \gamma_j + (1 - \gamma_j)[1 + \exp\{-\alpha_j(\theta - \beta_j)\}]^{-1}, \tag{1}$$

where α_j, β_j, and γ_j are the item discrimination, difficulty, and pseudo-guessing parameters, respectively, for item j, and θ is the ability parameter. Equation 1 expresses the probability of a correct answer on the θ scale. We will make use below of the fact that this probability is also the true score function for item j.

Estimation of the pseudo-guessing parameter is well-known to be problematic unless there is a large number of examinees for whom the item is reasonably difficult (cf. Baker, 1987, 1988; Kolen, 1981; Shepard, Camilli, & Averill, 1981; Thissen & Wainer, 1982). In this regard, Lord (1980) presented a DIF detection procedure which does not consider the pseudo-guessing parameter. The extension of Lord's chi-square procedure described in this paper likewise is discussed with the two-parameter logistic IRF defined as

$$P_j(\theta) = [1 + \exp\{-\alpha_j(\theta - \beta_j)\}]^{-1}. \tag{2}$$

The method, however, can be adapted easily to include comparison of the pseudo-guessing parameter and is sufficiently general to accommodate any IRT model for dichotomously scored items. Further, the method as described here can be applied to the Rasch logistic model (Rasch,1980) or extended to include Samejima's (1969) graded response model.

Definition of DIF

When item parameters are estimated from two different groups of examinees, we obtain two sets of item parameter estimates, $\left(a_{j1} \quad b_{j1} \right)$ from the first group and $\left(a_{j2} \quad b_{j2} \right)$ from the second. IRT assumes that the item parameters are invariant across groups if examinees are drawn from the same population (cf. Baker, 1985, 1992; Hambleton, 1989; Lord, 1980). Therefore, the two sets o: item parameter estimates should be identical within sampling fluctuation after proper scaling adjustment. When the parameter estimates in the first group are not the same as the estimates in the second, the item is considered to be functioning differentially in the two groups. Since the shapes of the IRFs are dictated by their parameters, when the parameters differ so will the IRFs.

The definition of DIF stated in terms of IRFs, however, is unnecessarily restrictive as it is applicable only for dichotomous scored items. A more useful definition of DIF would be the following:

> An item is considered to be functioning differentially when the item true score functions in the different groups are not equal.

As noted earlier, for the dichotomously scored items, the item true score function is identical to the IRF. The statement in terms of item true score functions provides a consistent definition of DIF to include not only dichotomous models but also polytomous IRT models in general (Cohen, Kim, & Baker, 1992). Further, it should be noted that the item true score functions are identical if and only if the sets of the item parameters from the groups are equal. Consequently, the null hypothesis for testing the equality of the two-parameter IRFs from K groups of examinees can be stated as

$$\mathrm{H}_0 : \begin{pmatrix} \alpha_{j1} \\ \beta_{j1} \end{pmatrix} = \cdots = \begin{pmatrix} \alpha_{jk} \\ \beta_{jk} \end{pmatrix} = \cdots = \begin{pmatrix} \alpha_{jK} \\ \beta_{jK} \end{pmatrix} . \tag{3}$$

The null hypothesis can also be stated as

$$\mathrm{H}_0 : \underset{\sim}{\xi}_{j1} = \cdots = \underset{\sim}{\xi}_{jk} = \cdots = \underset{\sim}{\xi}_{jK}, \tag{4}$$

where $\underset{\sim}{\xi}_{jk} = \begin{pmatrix} \alpha_{jk} & \beta_{jk} \end{pmatrix}'$. The alternative hypothesis is, of course,

$$\mathrm{H}_1 : \mathrm{H}_0 \text{ is not true.} \tag{5}$$

Lord's Chi-Square

For two groups of examinees, Lord (1980) presented a chi-square method for comparing vectors of item parameters. Lord's chi-square is obtained as follows: Suppose we define $\mathbf{v}_{jk}$ and $\underset{\sim}{\Sigma}_{jk}$ as the vector of maximum likelihood item parameter estimators and the asymptotic variance and covariance matrix for $\mathbf{v}_{jk}$, respectively, for the kth group of examinees. That is,

$$\mathbf{v}_{jk} = \begin{pmatrix} a_{jk} & b_{jk} \end{pmatrix}' \tag{6}$$

and

$$\underset{\sim}{\Sigma}_{jk} = \begin{pmatrix} \mathrm{var}(a_{jk}) & \mathrm{cov}(a_{jk}, b_{jk}) \\ \mathrm{cov}(a_{jk}, b_{jk}) & \mathrm{var}(b_{jk}) \end{pmatrix} . \tag{7}$$

Then, for large samples

$$\mathbf{v}_{jk} - \underline{\xi}_{jk} \sim N(0, \underline{\Sigma}_{jk}) \tag{8}$$

or equivalently

$$(\mathbf{v}_{jk} - \underline{\xi}_{jk})' \underline{\Sigma}_{jk}^{-1} (\mathbf{v}_{jk} - \underline{\xi}_{jk}) \sim \chi^2_{df=2}. \tag{9}$$

The statistic in Equation 9 is sometimes called the Wald statistic and can be used to make inferences about $\underline{\xi}_{jk}$ (Rubin, 1988; Wald, 1943). Hence, item parameters estimated in two groups of examinees, which have been placed on the same scale, can be compared using the following chi-square test statistic described by Lord (1980):

$$\chi^2_j = (\mathbf{v}_{j1} - \mathbf{v}_{j2})' (\underline{\Sigma}_{j1} + \underline{\Sigma}_{j2})^{-1} (\mathbf{v}_{j1} - \mathbf{v}_{j2}). \tag{10}$$

The null hypothesis tested is $\mathrm{H}_0 : \underline{\xi}_{j1} = \underline{\xi}_{j2}$. Lord's chi-square has two degrees of freedom for the two-parameter model. This chi-square has been shown to be effective for detection of DIF (Candell & Hulin, 1987; McCauley & Mendoza, 1985) and is based on the following assumption (Lord, 1980):

1. It is asymptotic.
2. θ are assumed to be known.
3. It is appropriate only for maximum likelihood estimates.

A DIF Statistic in Multiple Groups

Suppose we have a set of item parameter estimates for the two-parameter model for item j from K different groups of examinees. We assume that all item parameter estimates are placed on the same metric so that the comparisons can be made. We define $\mathbf{v}_j$ as the vector of length $2K$ of estimators of all item parameters from the K different groups, $\underline{\xi}_j$ as the vector of item parameters, and $\underline{\Sigma}_j$ as the block-diagonal and non-singular dispersion matrix of $\mathbf{v}_j$. The first step is to formulate the following model to describe $\mathbf{v}_j$ as

$$\mathbf{v}_j = \mathbf{X}\underline{\xi}_j + \underline{\varepsilon}_j, \tag{11}$$

where

$$\mathbf{v}_j = \begin{pmatrix} a_{j1} & b_{j1} & \cdots & a_{jK} & b_{jK} \end{pmatrix}', \tag{12}$$

$\mathbf{X}$ is a known design matrix such as $\mathbf{X} = \mathbf{I}_{2K}$,

$$\underline{\xi}_j = \begin{pmatrix} \alpha_{j1} & \beta_{j1} & \cdots & \alpha_{jK} & \beta_{jK} \end{pmatrix}', \tag{13}$$

and $\underline{\varepsilon}_j$ is the error vector with the dispersion matrix $D(\underline{\varepsilon}_j) = \underline{\Sigma}_j$ as

$$\underline{\Sigma}_j = \begin{pmatrix} \mathrm{var}(a_{j1}) & \mathrm{cov}(a_{j1}, b_{j1}) & \cdots & 0 & 0 \\ \mathrm{cov}(a_{j1}, b_{j1}) & \mathrm{var}(b_{j1}) & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & \mathrm{var}(a_{jK}) & \mathrm{cov}(a_{jK}, b_{jK}) \\ 0 & 0 & \cdots & \mathrm{cov}(a_{jK}, b_{jK}) & \mathrm{var}(b_{jK}) \end{pmatrix}. \tag{14}$$

It can be seen that, asymptotically,

$$(\mathbf{v}_j - \underline{\xi}_j)'\underline{\Sigma}_j^{-1}(\mathbf{v}_j - \underline{\xi}_j) \sim \chi^2_{2K} \tag{15}$$

provided that $E(\mathbf{v}_j) = \underline{\xi}_j$ (Dobson, 1992). We can re-express Equation 15 in terms of a new parameter vector $\mathbf{C}\underline{\xi}_j$ of length p such that the new variance and covariance matrix $\mathbf{C}\underline{\Sigma}_j\mathbf{C}'$ is non-singular. Here, $\mathbf{C}$ is a contrast matrix which contains p rows of contrast vectors that are linearly independent

(Johnson & Wichern, 1992). Then the quadratic term Q_j is defined as

$$Q_j = (\mathbf{C}\mathbf{v}_j - \mathbf{C}\underline{\xi}_j)'(\mathbf{C}\underline{\Sigma}_j\mathbf{C}')^{-1}(\mathbf{C}\mathbf{v}_j - \mathbf{C}\underline{\xi}_j). \qquad (16)$$

Any test for the homogeneity of item parameters can be expressed as

$$\mathrm{H}_0 : \mathbf{C}\underline{\xi}_j = \underline{0}. \qquad (17)$$

The asymptotic distribution of the quadratic term Q_j, which is a multi-group DIF statistic, under the null hypothesis, $\mathbf{C}\underline{\xi}_j = \underline{0}$, is given by

$$Q_j = (\mathbf{C}\mathbf{v}_j)'(\mathbf{C}\underline{\Sigma}_j\mathbf{C}')^{-1}(\mathbf{C}\mathbf{v}_j) \sim \chi^2_p, \qquad (18)$$

where p is the rank of $\mathbf{C}$ (Dobson, 1990).

For example, when we have two groups of examinees, we obtain two sets of item parameter estimates. We assume that a proper scaling adjustment has been done so that two item parameter estimates from the first and second groups are expressed on the same scale. Then,

$$\mathbf{v}_j = \begin{pmatrix} a_{j1} & b_{j1} & a_{j2} & b_{j2} \end{pmatrix}' \qquad (19)$$

and

$$\underline{\Sigma}_j = \begin{pmatrix} \mathrm{var}(a_{j1}) & \mathrm{cov}(a_{j1}, b_{j1}) & 0 & 0 \\ \mathrm{cov}(a_{j1}, b_{j1}) & \mathrm{var}(b_{j1}) & 0 & 0 \\ 0 & 0 & \mathrm{var}(a_{j2}) & \mathrm{cov}(a_{j2}, b_{j2}) \\ 0 & 0 & \mathrm{cov}(a_{j2}, b_{j2}) & \mathrm{var}(b_{j2}) \end{pmatrix}. \qquad (20)$$

The hypothesis of the equality of two sets of item parameters can be tested with the matrix of contrast coefficients defined as

$$\mathbf{C} = \begin{pmatrix} 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & -1 \end{pmatrix} \qquad (21)$$

which has rank of two. This yields

$$\mathbf{C}\underline{\xi}_j = \begin{pmatrix} 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & -1 \end{pmatrix} \begin{pmatrix} \alpha_{j1} \\ \beta_{j1} \\ \alpha_{j2} \\ \beta_{j2} \end{pmatrix} = \begin{pmatrix} \alpha_{j1} - \alpha_{j2} \\ \beta_{j1} - \beta_{j2} \end{pmatrix} \tag{22}$$

and

$$\mathbf{C}\mathbf{v}_j = \begin{pmatrix} a_{j1} - a_{j2} \\ b_{j1} - b_{j2} \end{pmatrix}. \tag{23}$$

With the null hypothesis $\mathbf{C}\underline{\xi}_j = \begin{pmatrix} 0 & 0 \end{pmatrix}'$, the test statistic can be written as

$$Q_j = (\mathbf{v}_{j1} - \mathbf{v}_{j2})'(\underline{\Sigma}_{j1} + \underline{\Sigma}_{j2})^{-1}(\mathbf{v}_{j1} - \mathbf{v}_{j2}) \tag{24}$$

which is in fact the same as Lord's chi-square in Equation 10 with two degrees of freedom.

Example

We provide an example to illustrate the detection of DIF in three groups of examinees using the method described above. The data for this example were taken from a study by Cohen and Kim (1992) of calculator effects on mathematics test items.

Data and Item Parameter Estimation

Three groups of 200 students each were selected from students enrolled in calculus and pre-calculus mathematics courses in Fall 1990 at a large midwestern university. The first group was composed of examinees who were not allowed to use calculators during the test (i.e., No-Calculator Group). The second and third groups were composed of examinees who used two different brands of scientific calculators when they took the test (i.e., Calculator-1 and Calculator-2 Groups). All students were tested during the first week of classes, prior to any instruction in the course.

The test consisted of 14 items assembled from the pre-calculus section of a standardized multiple-choice university mathematics placement test with five options per item. The items used were all operational items on the test and had originally been written for use without calculators. DIF analysis was used to determine which items were sensitive to calculator usage. It was also of interest to determine the effect of two different brands of calculators on item performance.

As the number of examinees in each group was relatively small, we opted to try to use the two-parameter logistic model to fit the data sets. IRT item parameter estimates were obtained using BILOG 3 (Mislevy & Bock, 1990) with the marginal maximum likelihood estimation. The item fit statistics provided by BILOG 3 indicated that the two-parameter model provided a good fit to the data. Use of an IRT model also assumes that the data are unidimensional. Reckase (1979) has suggested that this condition may be satisfied if the first component in a principal component analysis accounts for at least 20 percent of the variance. A principal component analysis using tetrachoric correlation coefficients indicated that the data sets were sufficiently unidimensional for purposes of this study.

Summary statistics for each group of examinees are given in Table 1. Examination of mean item difficulties for the test indicated that the test was at about the appropriate level of difficulty for the examinees in the sample. Classical test item difficulties and item-excluded biserial correlations are presented in Table 2. IRT item discrimination and item difficulty estimates along with estimated variances and covariances are given in Table 3.

Insert Tables 1, 2, and 3 about here

Iterative Linking

Under the assumption of item parameter invariance, item parameters estimated in different groups will differ from one group to another due only to errors in measurement. The metrics from these groups must first be equated to a common scale before between-groups DIF comparisons of parameter estimates are made. In the multi-group DIF study context, estimates from the calibration of the two focal (i.e., Calculator-1 and Calculator-2) groups must be transformed to the metric of the reference (i.e., No-Calculator) group. In this study, therefore, two sets of linear coefficients are required for transforming the estimates from each of the two focal groups to the reference group scale. For simplicity, we designate the No-Calculator group as the first group and the Calculator-1 and Calculator-2 groups as the second and third groups, respectively. To illustrate the transformation, the transformed estimates of item discrimination and item difficulty parameters from the second group to the metric of the first group for item j are given by

$$a_{j2}^{*} = a_{j2}/A \tag{25}$$

and

$$b_{j2}^{*} = Ab_{j2} + B, \tag{26}$$

where * indicates a transformed value and A and B are the linear transformation coefficients for linking. The task of linking two metrics is to determine appropriate coefficients A and B. Note that a different set of A and B coefficients needed to be determined to transform the third group estimates onto the scale of the first group. The test characteristic curve method by Stocking and Lord (1983) was used in this study as implemented in the computer program EQUATE (Baker, Al-Karni, & Al-Dosary, 1991) for determining the appropriate A and B coefficients.

The linking procedure may be seriously affected by the presence of DIF items (Lautenschlager & Park, 1988; Shepard, Camilli, & Williams, 1984). Therefore, an iterative linking procedure described by Candell and Drasgow (1988) was used in the DIF analyses in this paper. In this procedure, an initial set of linear coefficients is determined and used to transform the parameter estimates from the focal group to the reference group metric. DIF analyses are done and items identified as DIF items are removed and the linear coefficients re-calculated from the remaining items. Item parameter estimates from the focal group are again transformed onto the reference group metric and DIF analyses again conducted. This process continues until either no DIF items are detected or until the same set of DIF items is detected. Evidence suggests using the test characteristic curve method with iterative linking provided more accurate detection of DIF than either the weighted mean and sigma method or minimum chi-square method (Kim & Cohen, 1992).

DIF Measures

The null hypothesis used in the present study of the equality of three sets of item parameters was tested with the following contrast matrix:

$$\mathbf{C} = \begin{pmatrix} 1 & 0 & -1 & 0 & 0 & 0 \\ 0 & 1 & 0 & -1 & 0 & 0 \\ 1 & 0 & 0 & 0 & -1 & 0 \\ 0 & 1 & 0 & 0 & 0 & -1 \end{pmatrix}. \tag{27}$$

This contrast matrix is of rank four and yields the following comparisons:

$$\mathbf{C}\mathbf{v}_j = \begin{pmatrix} a_{j1} - a^*_{j2} \\ b_{j1} - b^*_{j2} \\ a_{j1} - a^*_{j3} \\ b_{j1} - b^*_{j3} \end{pmatrix}. \tag{28}$$

The null hypothesis for these comparisons is

$$\mathbf{C}\underline{\xi}_j = \begin{pmatrix} \alpha_{j1} - \alpha^*_{j2} \\ \beta_{j1} - \beta^*_{j2} \\ \alpha_{j1} - \alpha^*_{j3} \\ \beta_{j1} - \beta^*_{j3} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}. \tag{29}$$

The test statistics, Q_j, with four degrees of freedom can be obtained using Equation 18. In the present study Q_j was tested with a .05 type I error rate.

It is of course possible to specify a different contrast matrix than that given in Equation 27. For example, we could use

$$\mathbf{C} = \begin{pmatrix} 1 & 0 & -1 & 0 & 0 & 0 \\ 0 & 1 & 0 & -1 & 0 & 0 \\ 0 & 0 & 1 & 0 & -1 & 0 \\ 0 & 0 & 0 & 1 & 0 & -1 \end{pmatrix}. \tag{30}$$

This contrast matrix looks different than that in Equation 27 but would produce the same value of Q_j.

It is important to note that the calculation of Q_j requires that both item parameter estimates and the variance and covariance matrices from the second and third groups be placed onto the metric of the first group. As an example, the following transformations of the estimated variance terms from the second group are required in the calculation of Q_j:

$$\mathrm{var}(a^*_{j2}) = \mathrm{var}(a_{j2})/A^2 \tag{31}$$

and

$$\mathrm{var}(b^*_{j2}) = A^2\mathrm{var}(b_{j2}). \tag{32}$$

Note that no transformation is needed for the individual covariance terms.

For comparison purposes, three Lord's chi-squares were obtained from the pairs of the groups. Iterative linking was also used with Lord's chi-square for each pair of the groups. Since a .05 type I error rate was used for Q_j, each Lord's chi-square was tested with $1 - .95^{1/3} = .017$ type I error rate (Kirk, 1982).

As a comparison between the multi-group DIF statistic and Lord's chi-squares, three pairwise multi-group DIF statistics were also obtained. Estimates for this set of comparisons were based on the equating coefficients from the final iteration of the multi-group DIF Q_j procedure. These pairwise comparisons were also tested with a .017 type I error rate. For example, the pairwise comparison between the first and second group was obtained the contrast matrix defined as

$$\mathbf{C} = \begin{pmatrix} 1 & 0 & -1 & 0 & 0 & 0 \\ 0 & 1 & 0 & -1 & 0 & 0 \end{pmatrix}. \tag{33}$$

This pairwise Q_j may be the same as Lord's chi-square for the two groups compared. Differences which occur between the pairwise Q_j and Lord's chi-square do so because of the differences in linking coefficients obtained for the two approaches.

Results

Results for the multi-group Q_j, Lord's chi-square, and the pairwise Q_j are given in Table 4. Possible calculator effects were detected in two items (Item 10 and Item 14) using multi-group DIF statistic Q_j and in one item (Item 10) using Lord chi-square. The pairwise Q_j resulted in the same two DIF items (see Table 4) as the multi-group Q_j. Both Lord's chi-square and pairwise Q_j detected no significant differences on item performance between the two brands of calculators.

Insert Table 4 about here

The DIF detection procedure based on the multi-group Q_j required three linking iterations. (Recall that the pairwise Q_j were obtained from the transformed estimates in the final iteration.) Linkings based on Lord chi-square required one or two iterations. Table 5 contains linking coefficients and DIF items detected. Differences in the A and B coefficients after the first iteration are a result of the differences in the detection of DIF b`tween the multi-group DIF statistic and Lord's chi-square.

Insert Table 5 about here

The two items detected with Q_j were computational items and were easier when calculators were used. Item 10 was identified as a DIF item by the multi-group and pairwise DIF statistics. This item required examinees to find an unknown angle given the sine of a second unknown angle minus a known angle. This problem can be solved easily using a scientific calculator by entering each of the five choices and pressing the sine function. Examinees with calculators seemed to have an advantage on this item.

Item 14, a trigonometry item, was identified by both the multi-group and pairwise DIF statistics and Lord's chi-square. It required the examinee to find $\cos(-x)$ given the value of $\cos(x)$. Examinees with calculators had an advantage on this item. Using a calculator, an examinee could possibly have inserted a value for x and pressed function keys until an answer was found which agreed with one of the choices for the item.

Discussion

The presence of DIF in a test is a serious problem affecting the validity of the item as well as of the entire test. The typical DIF study is conducted between two groups. It is important to note that situations arise in which comparisons among several groups may be desirable or necessary. In such cases, one approach might be to conduct multiple pairwise comparisons. It may be preferable, however, to conduct simultaneous comparisons among the groups. In this paper, we presented a statistic, Q_j for simultaneous detection of DIF in multiple groups. This statistic is closely related to Lord's chi-square and is based on the same set of assumptions.

One of the assumptions of Lord's chi-square is that the θ are known. In tbis regard, Lord and Wingersky (1985) presented sampling variances and covariances of parameter estimates in IRT when abilities are unknown under the joint maximum likelihood estimation context. When θ are unknown, McLaughlin and Drasgow (1987) have shown that the type I error rate of Lord's chi-square may be seriously violated for joint maximum likelihood estimates. The type I error rate of Lord's chi-square does not appear to be inflated, however, when marginal maximum likelihood estimates or marginal Bayesian estimates were used (Cohen & Kim, in press). A well-known result of marginalized solutions (cf. Drasgow, 1989; Mislevy & Stocking, 1990) is that improved estimates of item parameters are typically obtained over those from joint maximum likelihood estimation. This improvement was shown in spite of the fact that ability was also treated as unknown. Further research is needed on the null distributions of Q_j, particularly for short tests and small samples.

After obtaining a significant multi-group DIF statistic, it may be of interest to compare pairs of groups. The pairwise Q_j could be used for this purpose. As discussed earlier, the pairwise Q_j is identical to Lord's chi-square but, as found in the present study, may be based on different equating coefficients. The results of the example give some indication of differences in the two procedures for the detection of DIF. It also should be noted that comparisons using the multi-group DIF statistics are not limited to pairwise cases. Comparison between the No-Calculator group and a combined calculator group (the Calculator-1 and Calculator-2 groups) could have been done using a contrast matrix such as the following:

$$\mathbf{C} = \begin{pmatrix} 1 & 0 & -.5 & 0 & -.5 & 0 \\ 0 & 1 & 0 & -.5 & 0 & -.5 \end{pmatrix}. \tag{34}$$

It should be noted that the type I error rate must be adjusted for the total number of contrasts (cf. Kirk, 1982).

The example given in this paper was used to illustrate a situation in which a comparison of item parameters was desirable among more than two groups. Using the multiple group Q_j statistic described in this paper, it was possible to simultaneously compare item parameters in each of the three groups in the example. Lord's chi-squares and the pairwise Q_j were also presented and used to illustrate the differences between the multi-group approach and the Lord chi-square approach. Differences between two types of DIF detection methods occurred because of the differences in linking coefficients. Even though the same general iterative linking procedure was used for two approaches, a different rationale for the statistics yielded different equating coefficients and, consequently, different sets of DIF items.

In the example provided in this paper, equal sample sizes of 200 for the reference and the two focal groups were used to control the effect of sample size. It is likely that in most DIF studies equal sample sizes do not occur. Further, the ability distributions for the reference and the two focal groups were well matched to the distribution of item difficulties. The effects of inequalities on these factors were not addressed in this paper.

An alternative approach to testing DIF in multiple groups might be one suggested by Lord (1980) which employs a MANOVA approach with post hoc comparisons based on Roy's method (Kim & Cohen, 1993). One drawback to this procedure, however, is that the assumptions of this method are somewhat more difficulty to realize than those for the multi-group DIF statistic presented in this study.

References

Baker, F. B. (1985). *The basics of item response theory.* Portsmouth, NH: Heinemann.

Baker, F. B. (1987). Methodology review: Item parameter estimation under the one-, two-, three-parameter logistic models. *Applied Psychological Measurement, 11*, 111-141.

Baker, F. B. (1988). The item log-likelihood surface for two- and three-parameter item characteristic curve models. *Applied Psychological Measurement, 12*, 387-395.

Baker, F. B. (1992). *Item response theory: Parameter estimation techniques.* New York: Marcel Dekker.

Baker, F. B., Al-Karni, A., & Al-Dosary, I. M. (1991). EQUATE: A computer program for the test characteristic curve method of IRT equating. *Applied Psychological Measurement, 15*, 78.

Berk, R. A. (Ed.). (1982). *Handbook of methods for detecting test bias,* Baltimore, MD: The Johns Hopkins University Press.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick, *Statistical theories of mental test scores* (pp. 395-479). Reading, MA: Addison-Wesley.

Candell, G. L., & Drasgow, F. (1988). An iterative procedure for linking metrics and assessing item bias in item response theory. *Applied Psychological Measurement, 12*, 253-260.

Candell, G. L., & Hulin, C. L. (1987). Cross-language and cross-cultural comparisons in scale translations: Independent sources of information about item nonequivalence. *Journal of Cross-Cultural Psychology*, *17*, 417-440.

Cohen, A. S., & Kim, S.-H. (in press). A comparison of Lord's χ^2 and Raju's area measures on detection of DIF. *Applied Psychological Measurement*.

Cohen, A. S., & Kim, S.-H. (1992). *Detecting calculator effects on a standardized mathematics test*. Madison: University of Wisconsin–Madison, Center for Placement Testing.

Cohen, A. S., Kim, S.-H., & Baker, F. B. (1992). *Detection of differential item functioning in the graded response model*. Madison: University cf Wisconsin–Madison, Center for Placement Testing.

Dobson, A. J. (1990). *An introduction to generalized linear models*. New York: Chapman and Hall.

Draba, R. E. (1977). *The identification and interpretation of item bias* (Research Memorandum No. 25). Chicago, IL: University of Chicago, Department of Education, Statistical Laboratory.

Drasgow, F. (1989). An evaluation of marginal maximum likelihood estimation for the two-parameter logistic model. *Applied Psychological Measurement*, *13*, 77-90.

Hambleton, R. K. (1989). Principles and selected applications of item response theory. In R. L. Linn (Ed.), *Educational measurement* (3rd ed.) (pp. 147-200). New York: Macmillan.

Holland, P. W., & Wainer, H. (Eds.). (1993). *Differential item functioning.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Johnson, R. A., & Wichern, D. W. (1992). *Applied multivariate statistical analysis* (2nd ed.). Englewood Cliffs, NJ: Prentice Hall.

Kim, S.-H., & Cohen, A. S. (1991). A comparison of two area measures for detecting differential item functioning. *Applied Psychological Measurement, 15*, 269-278.

Kim, S.-H., & Cohen, A. S. (1992). Effects of linking methods on detection of DIF. *Journal of Educational Measurement, 29*, 51-66.

Kim, S.-H., & Cohen, A. S. (1993). *Detection of DIF in multiple groups using MANOVA approach.* Madison: University of Wisconsin–Madison, Center for Placement Testing.

Kirk, R. E. (1982). *Experimental Design: Procedures for the behavioral sciences* (2nd ed.). Belmont, CA: Brooks/Cole.

Kolen, M. J. (1981). Comparison of traditional and item response theory methods of equating tests. *Journal of Educational Measurement, 18*, 1-12.

Lautenschlager, G. J., & Park, D.-G. (1988). IRT item bias detection procedures: Issues of model misspecification, robustness, and parameter linking. *Applied Psychological Measurement, 12*, 365-376.

Linn, R. L., Levine, M. V., Hastings, C. N., & Wardrop, J. L. (1981). Item bias in a test of reading comprehension. *Applied Psychological Measurement*, *5*, 159-173.

Lord, F. M. (1977). A study of item bias, using item characteristic curve theory. In Y. H. Poortinga (Ed.), *Basic problems in cross-cultural psychology* (pp. 19-29). Amsterdam, The Netherlands: Swets and Zeitlinger B. V.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Lord, F. M., & Wingersky, M. (1985). Sampling variances and covariances of parameter estimates in item response theory. In D. J. Weiss (Ed.), *Proceedings of the 1982 Item Response Theory and Computerized Adaptive Testing Conference* (pp. 69-88). Minneapolis: University of Minnesota, Department of Psychology, Computerized Adaptive Testing Laboratory.

McCauley, C. D., & Mendoza, J. (1985). A simulation study of item bias using a two-parameter item response model. *Applied Psychological Measurement*, *9*, 389-400.

McLaughlin, M. E., & Drasgow, F. (1987). Lord's chi-square test of item bias with estimated and with known person parameters. *Applied Psychological Measurement*, *11*, 161-173.

Mislevy, R. J., & Bock, R. D. (1990). *BILOG 3: Item analysis and test scoring with binary logistic models* [Computer program]. Mooresville, IN: Scientific Software.

Mislevy, R. J., & Stocking, M. L. (1990). A comsumer's guide to LOGIST and BILOG. *Applied Psychological Measurement*, *13*, 57-75.

Pine, S. M. (1977). Applications of item characteristic curve theory to the problem of test bias. In D. J. Weiss (Ed.), *Applications of computerized adaptive testing: Proceedings of a symposium presented at the 18th annual convention of the Military Testing Association* (Research Rep. No. 77-1, pp. 37-43). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program.

Raju, N. S. (1988). The area between two item characteristic curves. *Psychometrika*, *53*, 495-502.

Raju, N. S. (1990). Determining the significance of estimated signed and unsigned areas between two item response functions. *Applied Psychological Measurement*, *14*, 197-207.

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests*. Chicago, IL: The University of Chicago Press. (Original work published 1960)

Reckase, M. D. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. *Journal of Educational Statistics*, *4*, 207-230.

Rubin, D. B. (1988). Discussion. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 241-256). Hillsdale, NJ: Lawrence Erlbaum Associates.

Rudner, L. M. (1977, April). *An approach to biased item identification using latent trait measurement theory*. Paper presented at the annual meeting of the American Educational Research Association, New York. (ERIC Document Reproduction Service No. ED 137 337)

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometric Monograph, 17*.

Shepard, L., Camilli, G., & Averill, M. (1981). Comparison of procedures for detecting test-item bias with both internal and external ability criteria. *Journal of Educational Statistics*, *6*, 317-375.

Shepard, L., Camilli, G., & Williams, D. M. (1984). Accounting for statistical artifacts in item bias research. *Journal of Educational Statistics*, *9*, 93-128.

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement*, 7, 201-210.

Thissen, D., Steinberg, L., & Wainer, H. (1988). Use of item response theory in the study of group differences in trace lines. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 147-169). Hillsdale, NJ: Lawrence Erlbaum Associates.

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In

F. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 67-113). Hillsdale, NJ: Lawrence Erlbaum Associates.

Thissen, D., & Wainer, H. (1982). Some standard errors in item response theory. *Psychometrika*, *47*, 397-412.

Wald, A. (1943). Tests of statistical hypotheses concerning several parameters when the number of observation is large. *Transactions of the American Mathematical Society*, *54*, 426-482.

Wainer, H. (1993). Model-based standardized measurement of an item's differential impact. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 123-135). Hillsdale, NJ: Lawrence Erlbaum Associates.

Wright, B. D., & Stone, M. H. (1979). *Best test design*. Chicago, IL: MESA Press.

TABLE 1

Raw Score Summary Statistics for the Data Sets

	Group		
Statistic	No-Calculator	Calculator-1	Calculator-2
Number of Items	14	14	14
Mean Score	9.24	9.71	9.56
Standard Deviation	3.02	2.95	2.83
Coefficient Alpha	.66	.62	.60
Number of Examinees	200	200	200

TABLE 2

Classical Item Difficulties and Item-Excluded Biserial Correlations for the Data Sets

	No Calculator Group		Calculator-1 Group		Calculator-2 Group	
Item	Difficulty	Correlation	Difficulty	Correlation	Difficulty	Correlation
1	.86	.62	.88	.59	.86	.52
2	.88	.39	.92	.29	.92	.49
3	.75	.33	.70	.32	.70	.26
4	.76	.31	.74	.33	.73	.25
5	.72	.55	.74	.41	.68	.48
6	.76	.49	.77	.37	.77	.45
7	.52	.53	.49	.42	.53	.39
8	.69	.42	.79	.39	.72	.30
9	.67	.51	.67	.66	.72	.28
10	.55	.44	.70	.61	.66	.64
11	.71	.51	.70	.32	.69	.40
12	.57	.53	.59	.73	.54	.45
13	.45	.43	.49	.57	.47	.39
14	.37	.46	.54	.47	.60	.52

TABLE 3

Item Parameter Estimates and Their Estimated Variances and Covariances for the Data Sets

	No-Calculator Group			Calculator-1 Group			Calculator-2 Group		
Item	*a* (var)	*b* (var)	(cov)	*a* (var)	*b* (var)	(cov)	*a* (var)	*b* (var)	(cov)
1	1.53 (.17)	-1.73 (.10)	(.10)	1.39 (.15)	-2.01 (.15)	(.13)	1.18 (.11)	-2.06 (.18)	(.12)
2	.84 (.09)	-2.77 (.68)	(.24)	.66 (.10)	-4.06(2.86)	(.51)	1.22 (.20)	-2.58 (.40)	(.25)
3	.63 (.04)	-1.95 (.37)	(.11)	.55 (.04)	-1.69 (.36)	(.10)	.45 (.04)	-2.02 (.80)	(.16)
4	.56 (.05)	-2.19 (.71)	(.18)	.61 (.05)	-1.91 (.42)	(.12)	.45 (.03)	-2.37 (.96)	(.17)
5	1.16 (.08)	-1.10 (.06)	(.05)	.80 (.05)	-1.54 (.18)	(.08)	.80 (.05)	-1.01 (.08)	(.04)
6	.98 (.06)	-1.44 (.11)	(.06)	.77 (.05)	-1.79 (.24)	(.09)	.91 (.07)	-1.55 (.14)	(.08)
7	1.12 (.06)	-.10 (.03)	(.00)	.86 (.05)	.09 (.04)	(.00)	.68 (.04)	-.20 (.07)	(.01)
8	.79 (.05)	-1.16 (.12)	(.05)	.78 (.07)	-1.92 (.33)	(.13)	.58 (.04)	-1.79 (.38)	(.11)
9	1.02 (.06)	-.86 (.06)	(.03)	1.52 (.10)	-.71 (.03)	(.02)	.50 (.03)	-1.98 (.58)	(.12)
10	.91 (.05)	-.28 (.05)	(.01)	1.39 (.09)	-.87 (.04)	(.03)	1.52 (.13)	-.66 (.03)	(.03)
11	1.01 (.07)	-1.13 (.09)	(.06)	.58 (.04)	-1.60 (.30)	(.09)	.74 (.04)	-1.22 (.13)	(.05)
12	1.11 (.08)	-.34 (.04)	(.02)	1.89 (.15)	-.34 (.02)	(.01)	.91 (.05)	-.19 (.05)	(.01)
13	.84 (.04)	.29 (.05)	(-.01)	1.28 (.09)	.04 (.03)	(.00)	.68 (.04)	.24 (.07)	(-.01)
14	.94 (.06)	.70(.05)	(-.03)	.88 (.05)	-.25 (.04)	(.01)	1.05 (.06)	-.51 (.04)	(.02)

TABLE 4

Multi-Group DIF Statistics and Lord's Chi-Squares

	Multi-Group	Lord's Chi-Square			Pairwise Q		
Item	DIF Q	C1 vs NC	C2 vs NC	C1 vs C2	C1 vs NC	C2 vs NC	C1 vs C2
1	.15	.04	.04	.15	.03	.03	.16
2	1.76	.35	1.27	1.16	.58	1.27	1.61
3	1.77	2.45	1.06	.13	1.50	1.06	.07
4	.42	.58	.23	.24	.30	.23	.07
5	1.72	.87	.85	.94	.49	.85	1.20
6	.79	.45	.34	.51	.12	.33	.92
7	2.04	2.46	.66	1.61	.93	.66	1.76
8	3.78	2.05	.56	1.58	3.45	.56	1.42
9	6.76	2.43	1.61	7.57	2.57	1.62	6.09
10	11.46*	5.74	8.10	.23	8.51**	8.10	.68
11	1.84	2.14	.27	.41	1.39	.27	.85
12	4.42	3.32	.18	4.61	3.81	.18	3.19
13	2.98	1.80	.36	2.49	2.69	.35	1.59
14	20.20*	5.94	17.81**	2.69	8.70**	17.80**	3.55

*Significant at .05 alpha level and the crossponding critical value is $\chi^2_4 = 9.49$.

**Significant at .017 alpha level and the crossponding critical value is $\chi^2_2 = 8.16$.

TABLE 5

Linking Coefficients and DIF Items on Each Iteration

Method	1st Iteration		2nd Iteration		3rd Iteration	
(Linking)	Coefficients	DIF Item	Coefficients	DIF Item	Coefficients	DIF Item
Multi-Group DIF		14		10, 14		10, 14
(C1 onto NC)	$A = .957$ $B = .196$		$A = .934$ $B = .114$		$A = .896$ $B = .040$	
(C2 onto NC)	$A = .865$ $B = .101$		$A = .827$ $B =-.016$		$A = .788$ $B =-.080$	
Lord's Chi-Square						
C1 vs NC		None				
(C1 onto NC)	$A = .957$ $B = .196$					
C2 vs NC		14		14		
(C2 onto NC)	$A = .865$ $B = .101$		$A = .827$ $B =-.016$			
C1 vs C2		None				
(C2 onto C1)	$A = .899$ $B =-.101$					